데이터 인프라가 AI 성공의 기반인 이유

2023년이 인공지능(AI)이 '과한 기대감에서 현실로 옮겨간 해'로 기억(영문자료)된다면, 2024년은 단순한 관심에서 더 나아가 실질적인 결과를 얻어야 하는 해입니다. AI 성공을 위해서는 강력하고 안정적인 데이터 인프라가 필요하다는 메시지는 이제 IT 리더들에게 더 분명하게 다가옵니다.

속도, GPU, 대규모 언어 모델(LLM), 투자수익률(ROI)에 대해 많은 논의가 이루어지고 있는 상황에서 "AI 이니셔티브의 성공을 위해 데이터 플랫폼이 필수적인 이유는 무엇인가요? 데이터 플랫폼 없이 AI를 시도하면 어떻게 될까요?"와 같은 질문이 생길 수 있습니다.

이 글에서는 AI 혁신을 추진하기 위해 데이터 인프라가 갖춰야 할 기술적 역량에 대해 자세히 살펴봅니다.

AI의 기반인 데이터 인프라

AI를 데이터 워크로드의 F1 자동차라고 상상해 보세요. AI 프로젝트를 위한 데이터는 레거시 시스템에 저장된 과거의 데이터가 아니라 동적이고 방대하며 이동 중인 데이터입니다. 따라서 데이터를 저장, 액세스, 관리하는 방식이 중요합니다.

따라서 진행을 방해하고 궁극적으로 혁신을 늦추는 병목 현상을 일으키지 않으면서 데이터를 관리, 저장, 보호 및 제공하는 것이 도전과제입니다. 이를 위해서는 다음이 필요합니다:

- 원활한 데이터 접근성. AI 모델을 지연 없이 학습, 재학습, 배포할 수 있도록 해야 합니다. 속도도 중요하지만 시간이 지날수록 통합의 중요성이 더욱 커질 것입니다.
- 확장성. 성능과 속도를 유지하면서 증가하는 데이터 볼륨을 지원하려면 데이터 인프라는 새로운 데이터 파이프라인을 관리하는 오케스트레이션 툴과 함께 GPU 또는 CPU 클러스터 전반에서 수평적으로 확장할 수 있어야 합니다.
- 보안. 데이터 스토리지 플랫폼은 AI 모델이 의존하는 데이터 세트의 안전성, 복구 가능성, 컴플라이언스를 보장하기 위해 여러 계층의 데이터 보호, 보안, 거버넌스를 처리해야 합니다.

이제 논의의 핵심이 데이터 수집에서 적합한 데이터 아키텍처 설계로 변화했습니다. 적절한 비용에 성능, 확장성, 안정성, 효율성을 갖춘 데이터 인프라를 구축하려면 미래 지향적인 설계가 필요합니다.

AI를 위한 통합 스토리지 플랫폼의 요소

스토리지 플랫폼은 리소스 관리 및 성능 요구 사항을 완벽하게 해결할 수 있어야 합니다. 왜 그럴까요? 높은 처리량, 높은 동시성, 레이턴시가 낮은 액세스에 최적화된 스토리지 플랫폼이 없으면 AI 워크로드의 속도가 느려지거나 성능 기대치를 충족하지 못할 수 있기 때문입니다.

AI 성공을 위한 스토리지 플랫폼은 다음을 일관되게 제공해야 합니다:

파이프라인 오케스트레이션, 고대역폭 네트워킹을 통한 원활한 데이터 접근성

솔루션은 무엇보다도 우선 복잡한 인프라를 통합하고 데이터 운영을 간소화해야 하며, 이는 AI 파이프라인의 기반이 됩니다.

NVMe 기반 플래시 스토리지 또는 계층형 스토리지 솔루션은 높은 IOPS(초당 입출력 작업 수)를 제공할 수 있습니다. 새로운 데이터가 수

집, 변환, 배포될 때 쿠버네티스와 같은 오케스트레이션 툴은 스토리지 계층과 컴퓨팅 클러스터 간의 데이터 이동을 자동화하여 새로운 학습 데이터에 대한 지속적인 실시간 액세스를 보장할 수 있습니다.

AI 워크로드의 성공 여부는 빠른 스토리지뿐만 아니라 스토리지 시스템, 컴퓨팅 클러스터, 엣지 위치 간의 데이터 흐름을 처리할 수 있는 고속 네트워킹에도 달려 있습니다. 인피니밴드(InfiniBand) 또는 원격 직접 메모리 액세스(RDMA) 네트워킹을 사용하면, CPU 리소스를 소모하지 않고도 GPU 또는 컴퓨팅 노드 간에 초고속 데이터 전송이 가능하므로 전체 AI 파이프라인을 최적화할 수 있습니다.

더 자세한 내용은 다음 글에서 알아보세요: "데이터 플랫폼을 통해 사일로를 제거하여 AI 파이프라인을 가속화하는 방법"

GPU 활용 극대화를 위한 높은 처리량

그래픽 처리 장치(GPU)의 활용도를 고려하세요. "2024년 규모별 AI 인프라 현황(영문자료)" 보고서에 따르면 AI 워크로드는 증가하고 있지만, GPU 활용도는 여전히 도전과제로 남아 있습니다. 기업들은 사용 가능한 GPU 리소스를 제대로 활용하지 못하고 있으며, 피크 시간대에 85% 이상의 활용률을 보이는 기업은 7%에 불과하다고 응답했습니다.

AI 학습 작업에는 종종 신속한 병렬 데이터 액세스가 필요할 수 있습니다. **NVMe-oF**와 같은 기술을 활용하면, 회전 디스크나 느린 SSD의 기존 병목현상 없이 데이터를 GPU 지원 컴퓨팅 클러스터로 직접 스트리밍할 수 있습니다. 데이터 전송이 원활하게 이루어지면, GPU 활용도를 극대화하여 결과적으로 인사이트를 얻는 시간을 단축할 수 있습니다.

GPU 활용을 최적화하기 위한 아키텍처 결정에 대해 자세히 알아보기 >>

현재와 미래의 요구 사항을 위한 확장성

학습 모델은 데이터 인프라에 상당한 부담을 줄 수 있습니다. 모델이 반복되고 개선됨에 따라 실시간 데이터 접근성 및 피드백 루프(feedback loop)에 대한 필요성이 증가하여 인프라에 더 많은 부담을 줍니다.

데이터 세트가 커지고 모델이 복잡해지며 실시간 추론의 필요성이 확대됨에 따라, 현재는 충분할지 모르지만 내일은 부족할 수 있습니다. 따라서, 확장성은 타협할 수 없는 문제입니다. 이를 위해서는 여러 노드에서 동시에 데이터 액세스를 지원하는 분산 파일 시스템을 통합해야 합니다. 오브젝트 스토리지 시스템은 대용량 비정형 데이터를 확장 가능한 방식으로 저장할 수 있기 때문에 매우 중요합니다.

스토리지 플랫폼은 AI 이니셔티브가 발전함에 따라 쉽게 확장할 수 있어야 하며, 지속적인 인프라 업그레이드 없이도 증가하는 워크로드를 관리할 수 있어야 합니다. **S3** 호환 오브젝트 스토리지는 클라우드나 온프레미스 환경 전반에서 호환성 문제없이 AI 모델이 데이터를 수집할 수 있도록 보장합니다.

동급 최고의 데이터 복원력

AI를 위한 안전한 데이터 인프라를 구축한다는 것은 AI 파이프라인의 모든 지점에서 리스크를 완화하는 동시에 컴플라이언스, 안정성 및 데이터 거버넌스를 보장하는 것입니다.

AI 워크로드에는 클라우드, 온프레미스, 엣지 컴퓨팅 노드 등 다양한 환경 간에 민감한 데이터를 이동하는 경우가 많으므로 전송 및 미사용 시 암호화가 핵심입니다. 데이터베이스, 데이터 레이크 또는 분산 파일 시스템에 저장된 데이터 세트는 **AES-256**과 같은 표준을 사용하여 암호화할 수 있으므로 무단 액세스가 발생하더라도 데이터를 읽을 수 없습니다. 고급 시스템은 하드웨어 기반 암호화를 활용하여 사용 중인 데이터를 표적으로 삼는 공격을 방지합니다.

학습 및 추론을 위한 중요한 데이터 세트도 고가용성 및 복구가 가능해야 합니다. 자동화된 정기 스냅샷을 갖춘 데이터 플랫폼은 복구 지점을 사용

할 수 있도록 보장하며, 계층화된 백업 아키텍처는 이벤트 발생 시 비즈니스 연속성을 보장할 수 있습니다. 소거 코딩 및 RAID 구성은 하드웨어 장애로부터 데이터를 보호하고 다운타임을 최소화하면서 데이터를 복구할 수 있도록 합니다.

지속 가능한 AI 성장을 위한 비용 효율성

AI는 본질적으로 실험적이며 반복적이기 때문에 데이터 및 성능 요구 사항, 용량 수요를 예측할 수 없어 향후 용량 예측이 거의 불가능에 가깝습니다. 너무 적게 구축하면 고가의 GPU, 데이터 사이언스, AI 개발 투자에 대한 성능 병목 현상이 발생하고, 오버프로비저닝으로 소중한 예산이 낭비될 수 있습니다.

서비스형 소비 모델을 제공하는 비용 효율적인 스토리지 플랫폼(영문자료)은 성능의 한계 없이 GPU 클러스터의 최대 요구 사항을 기반으로 적절한 성능을 보장할 수 있습니다. 이를 통해 리소스를 현명하게 할당하여, 운영 비용을 늘리지 않고도 성능을 최적화할 수 있습니다.

지속 가능한 미래를 위한 에너지 효율성

AI 워크로드는 에너지를 많이 소비하기 때문에 상당한 전력 소비로 이어져, 비용이 증가할 뿐만 아니라 지속 가능성 이니셔티브에도 영향을 미칠 수 있습니다. GPU를 추가하여 데이터센터의 성능 한계를 뛰어넘기 시작하면 전력 및 냉각 한계에 부딪힐 가능성이 높습니다. 에너지 효율적인 스토리지(영문자료) 플랫폼으로 네트워킹 및 컴퓨팅과 동시에 스토리지를 처리하여 에너지 수요가 최대치에 도달하지 않도록 하세요. 이렇게 하면 성능(영문자료)을 유지하면서 공간, 전력, 냉각을 덜 필요로 하므로, 환경 목표에 맞춰 AI를 구현하고 에너지 풋프린트를 줄일 수 있습니다.

기술 리더: 견고한 데이터 기반을 통한 미래 지향적인 AI 프로젝트

성능이 뛰어난 통합 데이터 플랫폼이 아닌 다른 곳에서 AI 워크로드를 실행하려고 하는 것은 교통 체증 속에서 F1 자동차를 운전하는 것과 같습니다. 많은 정지와 이동, 그리고 마력 낭비가 예상됩니다. 하지만 일관되고 AI에 최적화된 플랫폼에서는 모든 것이 순조롭게 진행됩니다.

퓨어스토리지의 AI-레디 인프라 살펴보기

퓨어스토리지의 에이리(AIRI) AI-레디 인프라는 이러한 모든 측면을 충족하며, 까다로운 AI 워크로드를 지원하도록 설계되었습니다. 퓨어스토리지의 올플래시 스토리지 솔루션은 미디어 유형에 관계없이 대규모 데이터 세트를 처리하도록 설계되었으며, AI 운영의 성장에 따라 확장할 수 있어, 기업이 원하는 속도에 맞춰 혁신할 수 있도록 지원합니다.

견고하고 확장 가능한 고성능 데이터 아키텍처를 구축하면 현재와 미래의 AI 수요를 모두 충족할 수 있습니다. AI 분야에서는 기반이 큰 차이를 만들기 때문에, 이러한 아키텍처를 구축하는 것이 핵심적인 역할을 합니다.